www.LearningHealth.org

eSource Symposium: Resolving a Weak Link in the Learning Health Cycle

Summary Proceedings

A major barrier to accelerating learning health cycles remains the difficulty in efficiently and directly accessing high quality health care (real world eSource) data directly for research, including regulated research, to support the entire learning health cycle. In August, the Learning Health Community continued to explore this topic through the **2020 eSource Symposium: Resolving a Weak Link in the Learning Health Cycle.**

This Symposium consisted of three Seminars, each including six presentations by experts in these areas, examining various aspects of eSource in depth:

- Current Perspectives on eSource, RWD and Regulations
- Technology, Data Standards and Harmonization to Support eSource
- Expanding Learning Through Systems, Standards and Research

The goal of this Symposium was to build upon recommendations from the eSource Discussion Groups from the 2019 Bridging Collaborative and Society of Clinical Data Management (SCDM) Annual Conference, in addition to output from 'HL7 FHIR Blazing the Path Forward for Research' and the TransCelerate eSource initiative and other relevant activities and organizations.

We sincerely hope that this Symposium contributes to progress in the area of eSource (i.e. electronic source records) by convening thought leaders from across the spectrum — all united by the desire to strengthen connections between global clinical research and health care to build effective Learning Health Systems for the benefit of all patients. By sharing the results of related initiatives, we can all coalesce across the existing silos and converge on collaborative solutions while addressing remaining barriers that stand in the way of future progress.

The Learning Health Community (LHC) is a non-profit, charitable volunteer organization, which was initiated in 2012 and incorporated in 2017. The LHC adheres to a set of Core Values that are attributed to Learning Health Systems (LHS). Numerous organizations have endorsed these Core Values. In addition, we are very grateful for the LHC Advocates who supported this Symposium through contributions of resources and funds (Eli Lilly, Elligo Health Research, Verily, Syneos and Saama).

This Symposium was intended to take place as an in-person workshop at the National Academies of Science, Engineering and Medicine in Washington DC in May 2020. Unfortunately, the pandemic precluded this, and the Symposium became a virtual event. However, in the spirit of NASEM, we have prepared a set of proceedings. These proceedings are intended to be a brief summary of the Symposium with one representative slide from each presentation. **The actual presentations by the experts are posted on the LHC website (www.learninghealth.org), under Initiatives – 2020 LHC eSource Symposium, along with links to background reading, references and speaker photos and bios.**

Recommendations for Addressing A Weak Link in the Learning Health Cycle

Prior to the LHC eSource Symposium, each expert presenter was requested to respond to one common question: *What recommendations do you have for addressing a weak link in the Learning Health Cycle--to more effectively use electronic health record (EHR) data for clinical research*? Many of the responses were multi-faceted and many were overlapping, if not essentially the same. A summary of the consolidated and anonymized recommendations is shown in the following table.

Table 1. Recommendations for Addressing a Weak Link in the Learning Health Cycle

---

#1 - Structured data, data standards, controlled terminology, standardized common data models (CDMs), along with data governance processes and tools and resources to help use these, including mappings, computable biomedical knowledge and standardized clinical assessments

- Rapid convergence and adoption of CDISC-CDASH, CDISC TA standards, biomedical concepts with HL7 FHIR resources
- Providing access to patients of their data using USCDI
- Broad consent, standard agreements for data sharing
- Clear, unambiguous meaning for all data

#2 - Better EHR data quality and interoperability; consistent and high quality data collection at the point of care (also data security and protection and incentives for higher quality data); acknowledge that EHRs have fundamental/core issues that must be addressed to achieve high quality data and the desired interoperability.

#3 - Inclusion of data from many sources, not only EHR data from patients, but other RWD, including from wearables and healthy individuals; bridging the gap between data from academia and community practices

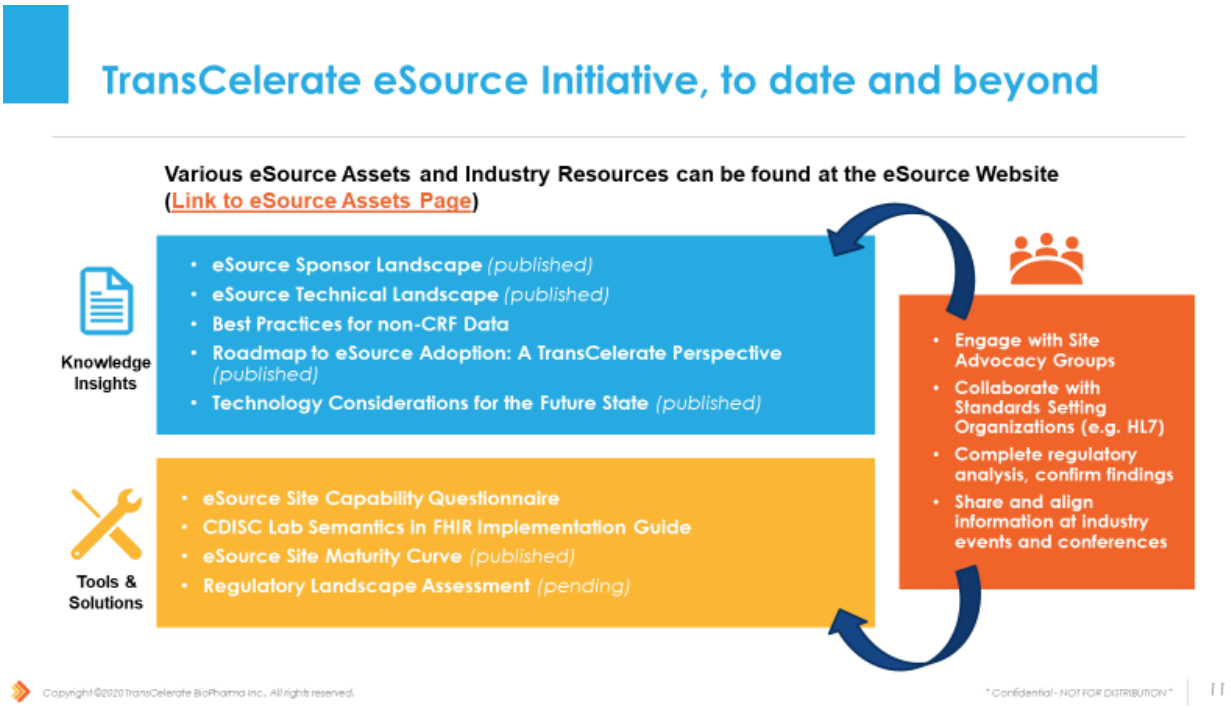
#4 - Collaboration (international) among ALL stakeholders, with a foundation that engenders trust and uses the learning/knowledge (data sharing), i.e. completes the learning health cycle.

- Put the patients first
- Provide education (e.g. to patients, IRBs) about data sharing
- Harmonize global regulatory and privacy expectations
- Reengineer healthcare processes

Table 1. Recommendations to address a weak link in the LHC—to more effectively use electronic health record (EHR) data for clinical research. The first two recommendations in this table were made by approximately half of the respondents (n=22, 19 expert speakers and 3 LHC leaders). The third and fourth recommendations were made by nearly one quarter of the respondents. In fifth place are four recommendations made by one respondent each.

Seminar 1: Current Perspectives on eSource, RWD and Regulations

- The first presenter in this seminar on Current Perspectives on eSource, RWD and Regulations was **Rakesh Maniar of Novartis**. He spoke on the topic of **TransCelerate's eSource Initiative**. Through this initiative, TransCelerate has produced a number of assets, which can be seen on the slide below.



TransCelerate has named four categories for eSource: Non-CRF data, Devices and Apps, Direct Data Capture and Electronic Health Records.

Mr. Maniar reported on the Regulatory Landscape Assessment, which is in progress. The team identified 17 key topics relevant to eSource after which they reviewed eSource guidance from various international regulators. From this assessment, they have identified where there is consensus and where there are differences or if a gap exists. The preliminary results are below.

Global State	Topic	Note
Gap	Archiving	EDC centric Consensus. Technology gap for eSource
Differences	Audit Trail	EDC centric guidance, ambiguity for eSource
Consensus	Control	Clear ICH defined expectations

Consensus	Data Integrity	Differences in level of details
Consensus	Data Privacy	Governed by local laws
Gap	Data Standards	No guidance
Differences	EHR and eSource Integration	No common guidance around integration
Gap	Fraud Detection	No guidance around eSource fraud detection
Consensus	Interim State	Clear on certified copy; difference in level of detail
Gap	Interoperability	Interoperability is encouraged yet no guidance on expectations
Consensus	Lineage/Traceability	Sufficient guidance exists
Gap	Monitoring Expectations	Paper and EDC centric vs eSource data volume
Gap	Ownership	Technology gap around apps and devices
Consensus	PI Oversight	Clear ICH defined expectations
Consensus	Source Definition	Clear definitions; lacks clarity around 3rd party storage
Consensus	Sponsor Oversight	Increased expectations of data transfer oversight
Differences	Validation	Specific for EHR

From this presentation, it is clear that we would benefit from harmonization and clarity across regulations related to eSource. Mr. Maniar also mentioned the Vulcan Accelerator and the desire for FHIR to be embraced by the health care industry and to support research to simplify implementation without sacrificing information integrity.

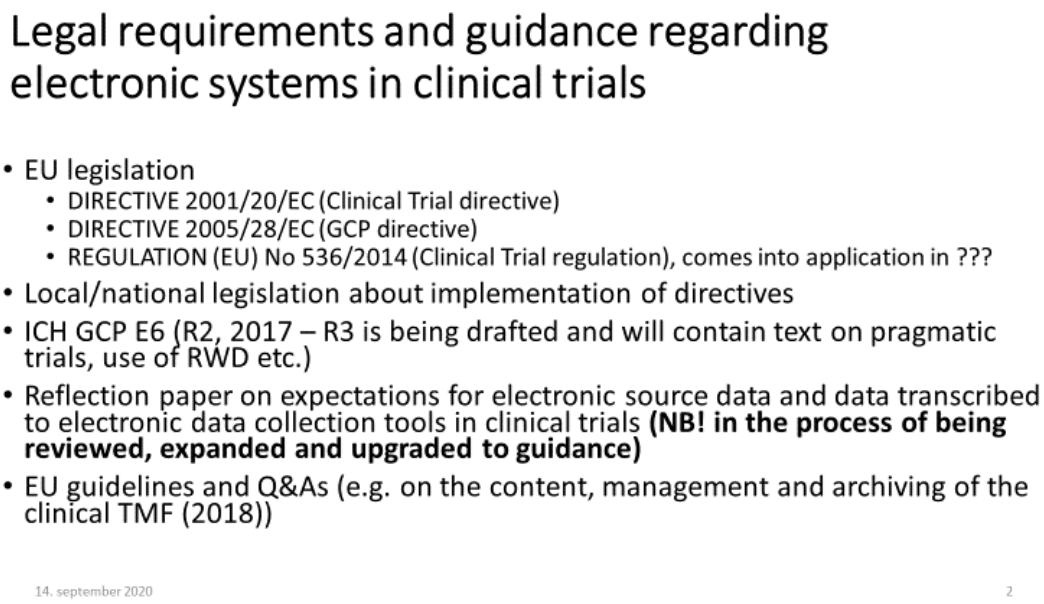
- The second speaker was **Lisbeth Bregnhøj of the Danish Medicines Agency, who is an EMA Good Clinical Practices Inspector.** She spoke about the **EMA Perspective on eSource, RWD and Opinion on eSource/DDC.**

Ms. Bregnhøj stated that GCP inspectors are interested in maintaining safety for the patients involved in clinical trials and that they ensure that the data are reliable. Critical deviations are "conditions, practices or processes that adversely affect the rights, safety or wellbeing of the subjects and/or the quality and integrity of data".

Typical areas of deviation related to eSource/DDC include data collection tools that 'lack fitness for purpose' due to the way they are designed or poor requirements; unclear definitions of source data;

lack of central metadata and data overview (including centralized monitoring); data protection breaches; inadequate IT security or user management; and, many others.

The relevant EMA/EU legal requirements and guidance are listed in this slide below.



Ms. Bregnhøj then provide four examples of inspection deviations identified during review of metadata:

a) Primary efficacy data that were not entered per protocol and were not 'true' source data
b) Audit trail review revealed that data entered in a patient diary (ePRO) could not have been entered in the timeframe required by the protocol
c) Access to the system was granted (and revoked) in a manner that compromised the blinding intended by the protocol (single-blind vs. double-blind)
d) User management was not conducted properly

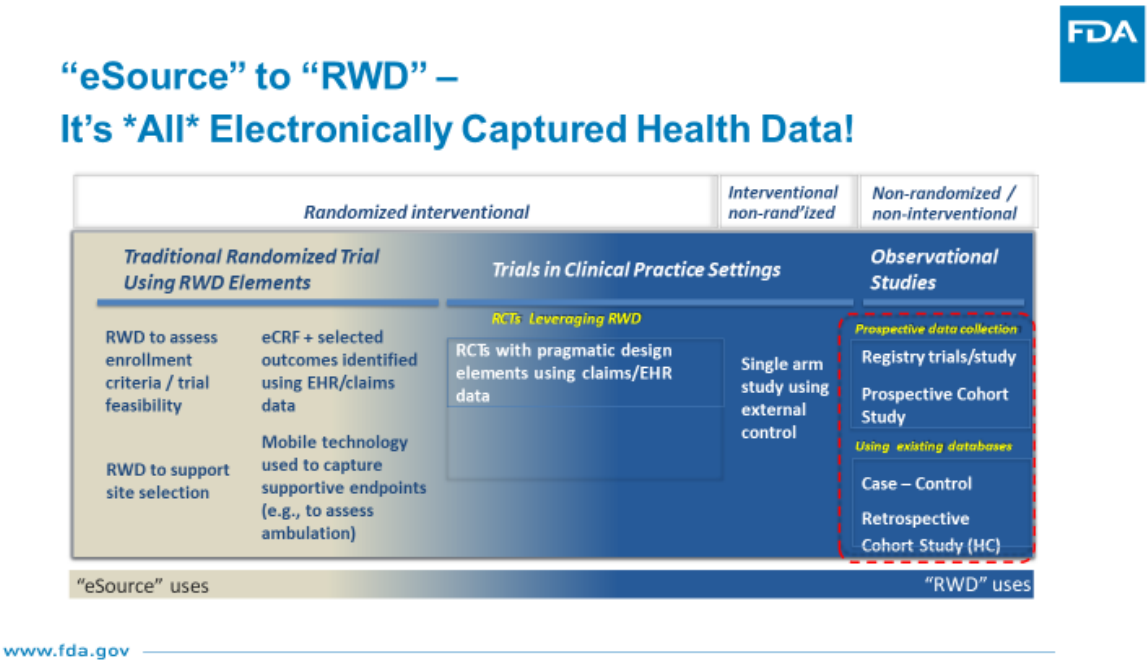
Important points about the review of metadata (e.g. audit trail, event logs) were made. GCP Inspectors expect that investigators have uninterrupted access to patient data, sponsors receive data and metadata from vendors/CROs on an ongoing basis, procedures for risk-based audit trail reviews are in place, and data are received in relevant formats.

❖ The third presentation in this seminar was given by **Mary Ann Slack of the U.S. Food and Drug Administration** on the topic of **FDA Guidance, Standards and Strategy on eSource and Real World Evidence.**

Encouragement for expanding the use of real world data within FDA has emanated from the 21st Century Cures Act of the U.S. Congress. A Real World Evidence Program was initiated that places increased focus on the use of RWD in clinical trials and/or to augment the data FDA received from RCTs. This program includes guidance, demonstration projects, stakeholder engagement and proposals for grants. Expanding RWD/E use "provides new opportunities to close the gap between research and clinical care". It provides 'big data', which facilitates the detection of infrequent outcomes or events and also enables a better understanding of how medications are actually used in practice. Ultimately, if

use and analysis of RWD/E can be perfected, it offers the potential to increase the scale of clinical research without a proportional increase in costs.

Ms. Slack defined eSource data as 'data captured electronically at its point of origin' and showed the following slide, which indicates the breadth of RWD sources.



Ms. Slack commented that currently there is limited FDA guidance related to RWD, such as the guidance on eSource and the use of EHRs and standardized data in clinical studies. Ms. Slack commented on what is currently 'underway' through this program, including a scan of relevant existing data standards, highlighting opportunities to improve or adapt existing standards for use and submission of RWD. The Objective, she stated, is "A Roadmap to a future state of standards that support RWD capture to regulatory submissions". Ms. Slack closed by stating that efforts in this space should be coordinated---a global culture of collaboration that builds on what has been done and does not reinvent the wheel.

- The fourth presentation in this Seminar was by **Jesper Kjær, who is the Director of the Danish Medicines Agency's Data Analytics Center**. He spoke about **"Personalized Medicine and True eSource".**

Mr. Kjær also started his presentation showing a slide of many different types of eSource data, including disease registries, social media, mortality data, survey data and other more common types of eSource such as wearables, lab data, EMR data, claims data and hospital and pharmacy data.

The vision of the Danish Medicines Agency's Data Analytics Center (DAC) is "Through use of clinical trial and real world data and advanced analytical methods we want to increase the accessibility of safe and effective medicines and medical devices."

By utilizing available data in new ways, the DKMA's DAC can help usher in a new regulatory paradigm. This is shown in the following slide.

DAC can help usher in a new regulatory paradigm

By utilizing the available data in new ways

Business area	Now	Future
Scientific advice	• Increasingly complex advanced clinical trial designs • No quantitative scientific advice	• Authorities expanded capacity → quantitative scientific advice • Better regulatory framework
Pre-approval	• Development and approval > 10 years • Approval based on RCT • Authorities review summaries of data	• Precision Medicine • Breakthrough / PRIME • Authorities have access to applicant's clinical trial data (known as CDISC formatted data) • Conditional approvals expands
Post-approval	• Subsequent detection of suspected side effects and signal generation	• Adverse drug reaction supplemented with real world data (RWD) • RWD: Registries, EHR, SoMe etc.

1

LÆGEMIDDELSTYRELSEN
DANISH MEDICINES AGENCY

Mr. Kjær then gave two examples of ongoing DAC activities. One is the platform that is shared with the Danish National Genome Center (NGC). Leveraging the NGC supercomputing power, they are able to ensure traceability of data use and reproducibility of analyses while controlling access to anonymized data and permissioned access for researchers. This enables the developments such as automated real-time AI/ML and its use in data analytics.

Another ongoing initiative is a collaboration with academia and other national authorities around the Covid-19 data. They have a list of research questions they are exploring, including the use of NSAIDs, anti-inflammatory pharmaceuticals and renin-angiotensin-aldosterone system inhibitors. (Status can be followed at http://www.lmst.dk/DACCOVID.)

DAC aims to be a trusted data broker using high quality data in a fully transparent and reproduceable manner, playing a direction-setting role in life science and an ethical sound use of big data and AI/ML in the life science ecosystem. This should help industry and patients to make drugs and technology faster and safely available. Ultimately, they wish to benefit patients by reducing inequality in health care with data analytics while protecting data.

Jesper Kjær closed by recommending awareness of the Priority Recommendations of the HMA-EMA (Heads of Medicines Agencies/European Medicines Agency) joint Big Data Task Force, which was established in 2017 to explore the use of big data to improve human and animal health. See the following.

Priority Recommendations of the HMA-EMA joint Big Data Task Force

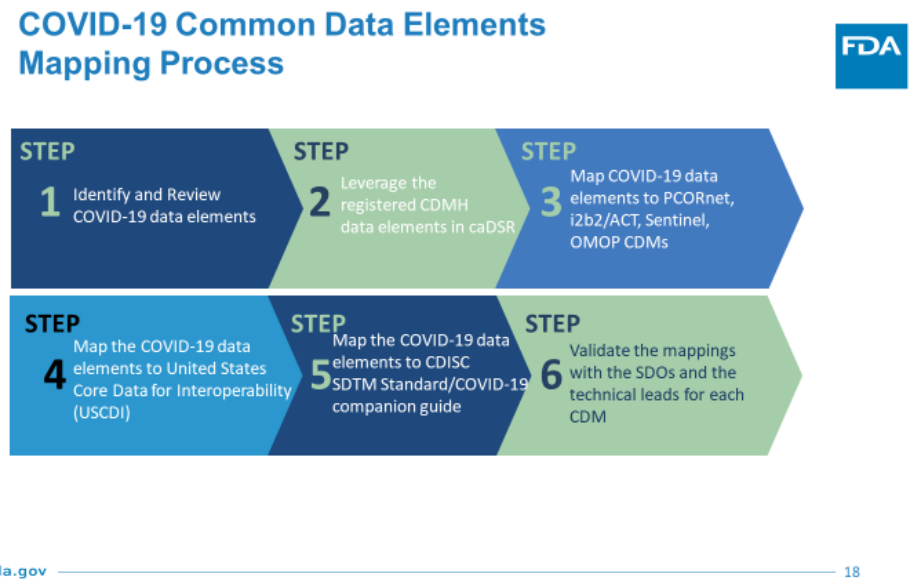
❖ The fifth speaker in this Seminar was **Mitra Rocca of FDA**, who spoke about **The Common Data Model Harmonization (CDMH) Project.**

Ms. Rocca stated that the goal of the CDMH project, which was funded through the Patient-Centered Outcomes Research (PCOR) Trust Fund, was "to build a data infrastructure for conducting research using Real World Data (RWD) derived from the delivery of health care in routine clinical settings". The specific objective was to "develop the method to harmonize the Common Data Models of various networks, allowing researchers to simply ask research questions on much larger amounts of RWD than currently possible, leveraging open standard and controlled terminologies to advance PCOR.

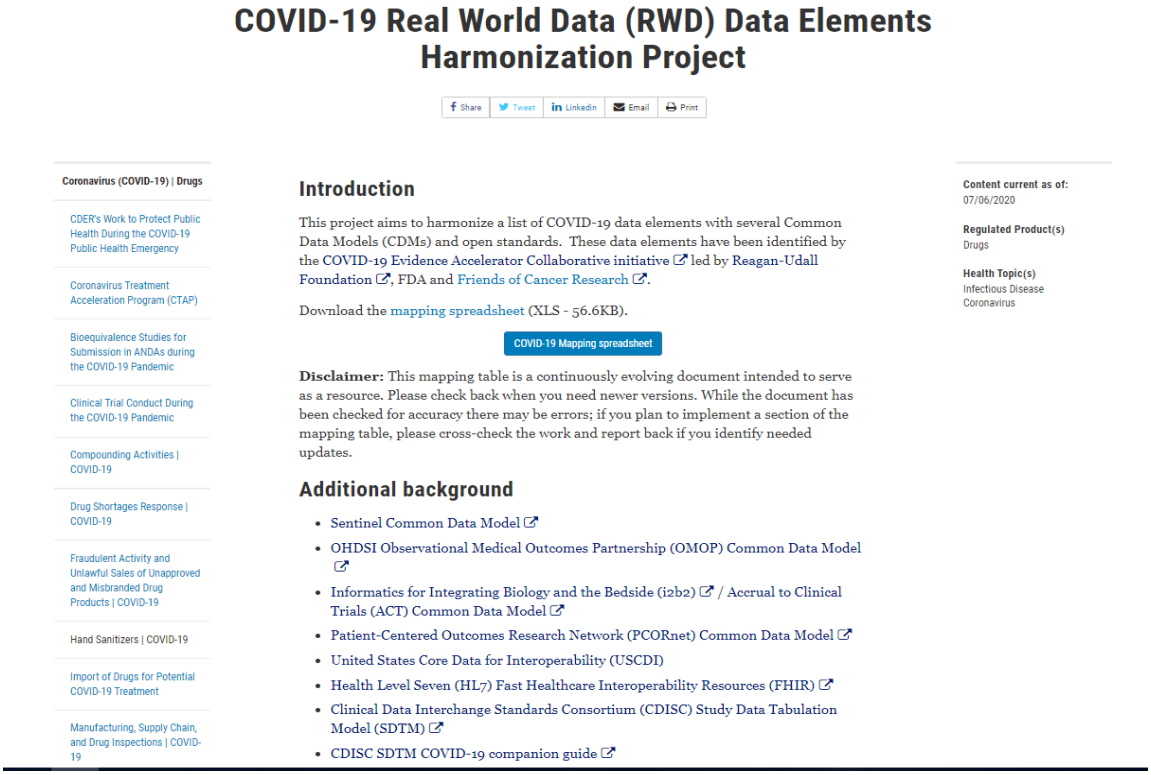
The current networks include Sentinel, i2b2/ACT, PCORNet, OMOP. The Phase I of this project has been completed. Accomplishments include harmonized common data models (CDMs) from the four networks with the Biomedical Integrated Domain Group (BRIDG) model, infrastructure development and collaborations with Yale/Mayo Clinic and Elligo Health Research to execute a query focused on an oncology (immunotherapy) use case. Phase II of the CDMH project is in progress. New data partners will be identified by Elligo Health Research to leverage the CDMH architecture; the existing architecture will be enhanced to leverage HL7 FHIR and the RWD will be submitted to FDA in CDISC SDTM format via the FDA Gateway.

Ms. Rocca also reported on three other related FDA activities that are now focused on Covid-19: a) the National COVID Cohort Collaborative (N3C), b) the Covid-19 Evidence Accelerator Collaborative and c) the I-SPY Covid-19 trial. The N3C is using the architecture from the CDMH project; however, they will use OMOP as the CDMH because most academic institutions involved in this project are not yet able to use HL7 FHIR. (More information about the N3C was presented by Dr. Ken Gersing for Seminar 3 of this Symposium.)

Ms. Rocca has done a significant amount of COVID-19 data elements mapping for the COVID-19 Evidence Accelerator project. See the slide below.



The mappings have been presented at the regular FDA Evidence Accelerator meetings led by Reagan-Udall Foundation and Friends of Cancer Research. (More information about this Accelerator was presented by Susan Winkler for Seminar 3 of this Symposium.) The mappings are also available through the FDA website.



COVID-19 Real World Data (RWD) Data Elements Harmonization Project

Share | Tweet | LinkedIn | Email | Print

Coronavirus (COVID-19) | Drugs

- CDER's Work to Protect Public Health During the COVID-19 Public Health Emergency
- Coronavirus Treatment Acceleration Program (CTAP)
- Bioequivalence Studies for Submission in ANDAs during the COVID-19 Pandemic
- Clinical Trial Conduct During the COVID-19 Pandemic
- Compounding Activities | COVID-19
- Drug Shortages Response | COVID-19
- Fraudulent Activity and Unlawful Sales of Unapproved and Misbranded Drug Products | COVID-19
- Hand Sanitizers | COVID-19
- Import of Drugs for Potential COVID-19 Treatment
- Manufacturing, Supply Chain, and Drug Inspections | COVID-19

Introduction

This project aims to harmonize a list of COVID-19 data elements with several Common Data Models (CDMs) and open standards. These data elements have been identified by the COVID-19 Evidence Accelerator Collaborative initiative led by Reagan-Udall Foundation, FDA and Friends of Cancer Research.

Download the mapping spreadsheet (XLS - 56.6KB).

COVID-19 Mapping spreadsheet

Disclaimer: This mapping table is a continuously evolving document intended to serve as a resource. Please check back when you need newer versions. While the document has been checked for accuracy there may be errors; if you plan to implement a section of the mapping table, please cross-check the work and report back if you identify needed updates.

Additional background

- Sentinel Common Data Model
- OHDSI Observational Medical Outcomes Partnership (OMOP) Common Data Model
- Informatics for Integrating Biology and the Bedside (i2b2) / Accrual to Clinical Trials (ACT) Common Data Model
- Patient-Centered Outcomes Research Network (PCORnet) Common Data Model
- United States Core Data for Interoperability (USCDI)
- Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR)
- Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM)
- CDISC SDTM COVID-19 companion guide

Content current as of: 07/06/2020

Regulated Product(s): Drugs

Health Topic(s): Infectious Disease, Coronavirus

https://www.fda.gov/drugs/coronavirus-covid-19-drugs/covid-19-real-world-data-rwd-data-elements-harmonization-project

In addition, Ms. Rocca is collaborating with the University of California in San Francisco on the I-SPY COVID-19 trial, which is leveraging the OneSource infrastructure that was build for the I-SPY-2 breast cancer trials. Goal is to minimize data entry, focus on clinical care and automate patient referrals and study registration with EHR integration. This trial is being conducted at UCSF.

- The sixth speaker of this Seminar was **Dr. Michael Ibara**, who is currently Head of Data Sciences at Elligo Health Research. He spoke about **Real World Experience with RWD in Community Practices.**

Dr. Ibara started this presentation by stating that, as GB Shaw might have said: *Healthcare and biopharma are two industries separated by a common set of data.* Requirements for biopharmaceutical researcher and the regulations to which they must adhere may seem arcane. However, the community researcher may be naïve in terms of these requirements, is typically understaffed (especially with respect to regulatory and technical expertise) and the economic incentives are not usually aligned to encourage them to do research. Healthcare data is often called 'dirty' or accused of being 'poor quality'. However, it is fit for care data (which does not mean it is free of issues such as completeness, accuracy or bias). That said, it is generally not fit for research as it stands. Electronic health records were developed with a focus on clinical concepts, scheduling and reimbursement. Regulated research focuses reproducibility, standards and statistics. To make such data 'research ready' requires informatics and

coding expertise and Data Science expertise, along with time and effort. The typical community practice physician has none of these.

Dr. Ibara then discussed three pillars of data for modern clinical research: clinical trial data, EHR data and patient-generated data. In the modernization of clinical trials, which is being encouraged by regulators, these data streams should all be taken into consideration and data collected from these sources simultaneously during trials. Efficiencies will arise from the ability to acquire, consume and analyze data across various sources. Also, there is a goal in trials now to allow the research subject to spend more time at home with minimum trips to the clinic or a physical site.

One approach to solving the problem that many physicians will not be able to participate in research is to provide an infrastructure to allow them to 'plug in' to research. This means that they need data, applications and a 'clinical control layer' that provides regulatory capabilities. Removing the need for special expertise and administrative overhead may provide an incentive for community participation in research. In other words, "to fix eSource, fix the underlying problems". We need to separate the data completely from its form factor and stop defining one in terms of the other. Think of data as the electricity of our times…it's not the machine, it's what powers the machine.

Seminar 2: Technology, Data Standards and Harmonization to Support eSource

- The first presenter in this seminar on Technology, Data Standards and Harmonization to Support eSource was **Denise Warzel** from the **National Cancer Institute.** Her presentation was entitled "CDMs, CDEs and BRIDG".

Ms. Warzel began by discussing FAIR data sharing (Findable, Accessible, Interoperable and Reusable) and defining Common Data Elements (CDEs) as discrete, clearly defined and reusable data collection Units (ISO/IEC 11179). One key issue with CDEs, however, is that they are not always 'common', not always well-constructed, and not usually unique.

There are two primary NIH repositories where CDEs are hosted. One is the NCI caDSR, which was established in 2000 and contains CDEs that are linked to controlled terminology curated through the NCI Enterprise Vocabulary Services. (This is where CDISC terminology is housed.) The other is the NIH CDE Repository, which was established in 2010 and contains CDEs from a subset of NIH Centers, primarily from NINDS. CDEs represent a step in the direction of enabling 'FAIR Data Sharing'; however, remaining barriers deter broader adoption. These include:

- lack of standardization and governance to harmonized and elevate specific CDEs to standards
- limited engagement in CDE development by the research community
- inconsistent enforcement efforts
- difficulty in implementation and use
- issues with harmonization or curation

Ms. Warzel then proceeded to discuss progress in the area of harmonization related to the CDMH project that was introduced on 20 August by Ms. Mitra Rocca, FDA. The NCI partnered with FDA for this project to a) map disparate CDMs (Sentinel, PCORNet, OHDSI/OMOP and i2b2) to the BRIDG model; b) extended BRIDG clinical content; c) ballot BRIDG with extensions through SDOs; d) register the CDMs as

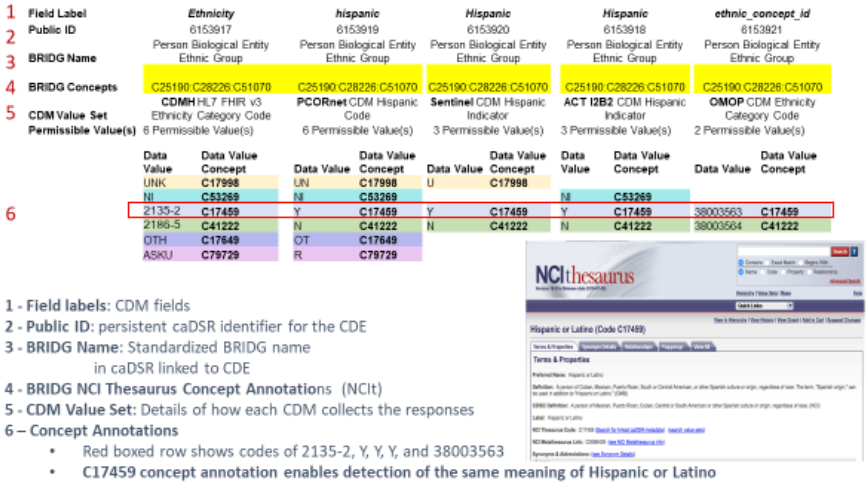
metadata in the caDSR database. The Metadata Annotation and Registration is very important in that metadata can make the meaning of data FAIR as follows.

- Findable: Data specification represented in human and machine-readable format
- Accessible: Metadata Repository with Web UI and API interfaces
- Interoperable: Normalizes the meaning of fields and the data values using standard terminology
- Reusable: Persistent, unique identifier including versions, semantic concept annotations supporting data validation and data transformation

Model	Version	Model Tables	Model Elements	Elements Found In BRIDG	Elements Not Found In BRIDG	Percent Elements Not Found In BRIDG	Elements Not Mapped
Sentinel	6.0.2	12	126	112	11	9%	3
PCORNET CDM	3.1	14	151	87	17	11%	46
OHDSI/OMOP	5.2	16	163	136	27	17%	*7 tbls*
i2b2	1.4	6	53	33	5	9%	15
Totals		**48**	**493**	**368**	**60**	**12%**	**64**

The BRIDG model was used as a "Semantic Hub" because as a CDISC, HL7 and ISO standard developed through a partnership with FDA, NCI, CDISC and HL7, it provides a conceptual basis for healthcare research data and as it is expanded to model more point-of-care healthcare concepts, it can serve as a bridge between this too related healthcare data originators. This model provides relationships and hierarchy between the concepts that are used in data models. The classes and attributes that are part of the BRIDG model are not bound to any coding system or terminology for data collection. However, the meaning of classes and attributes are annotated and registered as metadata using NCIt standard terminology concepts helping to provide unambiguous meaning for data. NCIt is the source of truth for CDISC and FDA terminology, linked to UMLS CUIs and is the preferred terminology for NCI. Phase II of the CDMH project will focus on using FHIR. An example of the semantic mapping among the four CDMs and BRIDG was provided. See the following slide illustrating how the data values for ethnicity are linked across different models using the NCIt concepts highlighted by the red outline.

caDSR Example: Semantic mapping of Hispanic/Ethnicity



	Field Label	Ethnicity	hispanic	Hispanic	Hispanic	ethnic_concept_id
2	Public ID	6153917	6153919	6153920	6153918	6153921
3	BRIDG Name	Person Biological Entity Ethnic Group	Person Biological Entity Ethnic Group	Person Biological Entity Ethnic Group	Person Biological Entity Ethnic Group	Person Biological Entity Ethnic Group
4	BRIDG Concepts	C25190:C28226:C51070	C25190:C28226:C51070	C25190:C28226:C51070	C25190:C28226:C51070	C25190:C28226:C51070
5	CDM Value Set Permissible Value(s)	**CDMH** HL7 FHIR v3 Ethnicity Category Code 6 Permissible Value(s)	**PCORnet** CDM Hispanic Code 6 Permissible Value(s)	**Sentinel** CDM Hispanic Indicator 3 Permissible Value(s)	**ACT I2B2** CDM Hispanic Indicator 3 Permissible Value(s)	**OMOP** CDM Ethnicity Category Code 2 Permissible Value(s)

Data Value	Data Value Concept	Data Value	Data Value Concept	Data Value	Data Value Concept	Data Value	Data Value Concept	Data Value	Data Value Concept
UNK	C17998	UN	C17998	U	C17998				
NI	C53269	NI	C53269			NI	C53269		
2135-2	C17459	Y	C17459	Y	C17459	Y	C17459	38003563	C17459
2186-5	C41222	N	C41222	N	C41222	N	C41222	38003564	C41222
OTH	C17649	OT	C17649						
ASKU	C79729	R	C79729						

1 - **Field labels**: CDM fields
2 - **Public ID**: persistent caDSR identifier for the CDE
3 - **BRIDG Name**: Standardized BRIDG name in caDSR linked to CDE
4 - **BRIDG NCI Thesaurus Concept Annotations** (NCIt)
5 - **CDM Value Set**: Details of how each CDM collects the responses
6 – **Concept Annotations**
- Red boxed row shows codes of 2135-2, Y, Y, Y, and 38003563
- **C17459 concept annotation enables detection of the same meaning of Hispanic or Latino**

NIH NATIONAL CANCER INSTITUTE 3

Ms. Warzel closed with Recommendations for FAIR Data Sharing as follows.

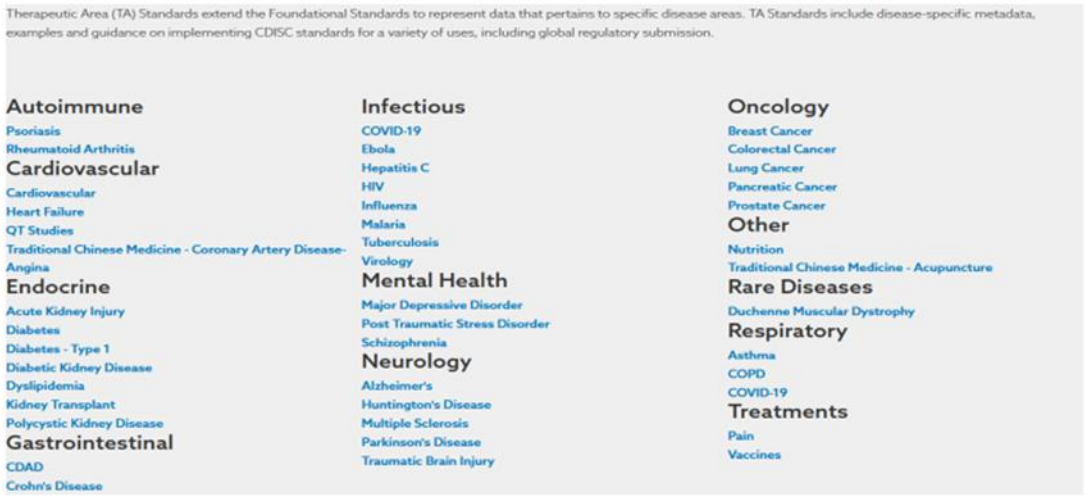
- Encourage the Acceptance and Re-use of Harmonized and Preferred CDEs and Global Data Standards
- Remove Political and Social Barriers to Data Sharing (which includes removing political barriers, aligning incentives and encouraging broad collaboration across multiple types of organizations; CDE governance should seize the opportunity to broaden CDE impact beyond niche implementation and by engaging the broader research and healthcare community
- Build Better 'Bridges' Between Research and Healthcare through mapped CDEs

"Overcoming these barriers is essential to efficiently and responsibly share meaningful data that can ultimately evolve into learning health systems through which research more rapidly informs care decisions for all of us as patients."
(See also the JBI reference on this topic, which is posted on the LHC website.)

❖ The second speaker in Seminar #2 was **Peter Van Reusel of CDISC**. He spoke about **Digitizing Therapeutic Areas: Increasing Standardization and Reusability.**

Mr. Van Reusel first recounted the history of the development of therapeutic area standards/user guides (TAUGs) within CDISC. These augment the CDISC foundational standards, which cover data that are common across all therapeutic areas. The first CDISC TAUG was developed in collaboration with the Critical Path Institute (C-Path) for Alzheimer's Disease. The Coalition of Accelerated Standards and Therapies (CFAST) was launched in 2012 with a broader group of collaborators, including FDA, TransCelerate, NIH/NCI, Innovative Medicines Initiative and ACRO, along with CDISC and C-Path. Over the past eight years, 44 TA standards have been developed following the robust CDISC standards development process. See the list below or on the CDISC website (www.cdisc.org).

Current Therapeutic Area User Guide Overview



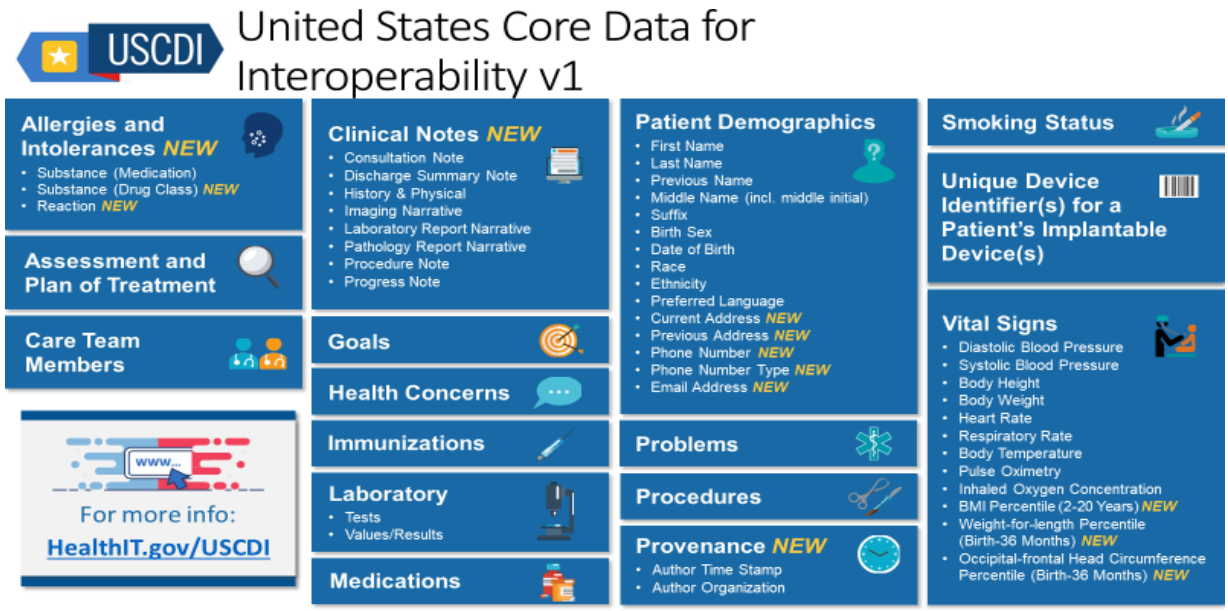
https://www.cdisc.org/standards/therapeutic-areas/disease-area

Mr. Van Reusel then went into more depth to describe that there is an initial concept mapping (scoping) step that facilitates communication between scientists and data standards experts. Case report forms are annotated and sample analysis datasets (results) are prepared so that the standards development can begin with 'the end in mind' and the data flow from collection to tabulation to analysis will be streamlined without requiring mapping and interpretation. The biomedical concepts that are developed create and store standards as concepts, thus creating meaning between data. They are electronically published as linked metadata and variability is minimized. A project called CDISC 360 will develop concept-based standards definitions, test and demonstrate end-to-end automation of study specification data processing and analysis. The TA standards should all evolve to support the end-to-end flow. In addition, CDISC is preparing a CDASH eCase Report Form (eCRF) Library to facilitate data collection in a standard format. Analysis Results Metadata is another area of work for the future.

- **Dr. Ida Sim, Professor of Medicine and Co-founder of Open mHealth and Vivli**, spoke next about the **HHS Office of the National Coordinator (ONC) Interoperability Rule: Implications for Access and Use of it US CORE.**

Dr. Sim reminded everyone that *interoperability* refers to the ability of computer systems to exchange information and make use of it. *Syntactic interoperability* is the format while *semantic interoperability* is the meaning. Examples of the former are exchange protocols over networks (e.g. HL7 FHIR, DICOM, XML) whereas the latter requires controlled vocabularies or codelists.

The U.S. ONC has announced in an Interoperability Rule that the syntax standard will be HL7 FHIR and the semantic standard will be the USCDI (US Core Data for Interoperability). Under this new Rule, all EHRs must offer an API that supports FHIR and USCDI. See the following slide for the contents of USCDI. Detailed information can be found at HealthIT.gov/USCDI.



In 2018, Apple Health Records brought USCDI to the iPhone. However, Android market share is 52% in the U.S. and 75% worldwide. Hence, in 2020, Common Health (with funding from Rockefeller Foundation) brought USCDI to Androids. This opportunity went live through Google Play store with UCSF as the first health system. CommonHealth was developed in partnership with UCSF, Cornell Tech, Sage Bionetworks and Open mHealth. The Commons Project is a 501c3 charitable non-profit public trust that was established to 'build digital services that put people first'. As Dr. Sim commented "The spigot is open" now. Patients can access their health data and also give permissions to use this in clinical research or for other purposes.

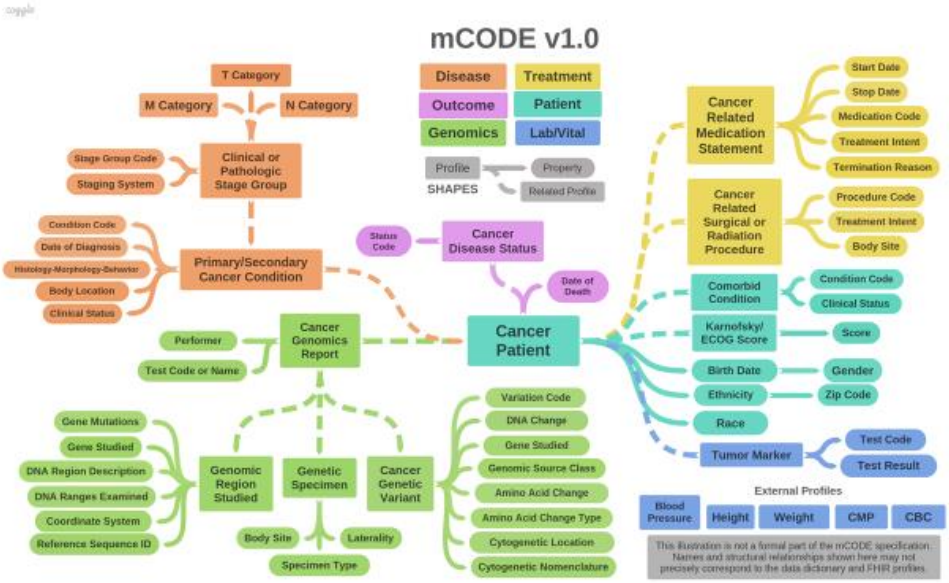
- **Dr. Monica Bertagnolli, a leader within Data-Farber/Brigham & Women's Cancer Center and the Alliance for Clinical Trials in Oncology** spoke on the topic of **mCODE and the CODEX Accelerator Experience.**

Dr. Bertagnolli is also the Past-Chair of the American Society of Clinical Oncology (ASCO) and was quite involved in Cancer LinQ. She provided an example of one key issue that was experienced through this data sharing initiative: within over 15.5 million entries and numerous EHRs, there were 51 different representations of Tobacco Use Assessment. See below for just a subset of these representations. This diversity in data entered/exchanged requires significant manual curation to make the data useful. The other option is that large amounts of potentially valuable data are lost when the exchanged information does not support 'semantic' interoperability.

Dr. Bertagnolli elaborated by indicating that diagnosis codes, encounter codes, infused medications, laboratory tests, physical exam data and smoking/pain assessments are generally available as structured data whereas other data are only sometimes available or generally NOT available as structured data, including oral medications, staging elements, disease status and surgery information. Perhaps the most important missing information for patients undergoing treatment is that indicating response to treatment. This deficit makes it impossible to adequately track one of the most important clinical outcomes, severely limiting the use of EHR-based datasets.

Value	Distinct Patients
Non-smoker	560,281
Never smoked tobacco	462,842
Ex-smoker	373,431
Current smoker	121,186
Unknown tobacco consumption	83,550
Smokes tobacco daily	81,250
Occasional tobacco smoker	22,607
Heavy smoker	5,898
Light tobacco smoker	3,478
Tobacco user	576
Current tobacco non-user	212
Chews tobacco	160
Passive smoker	140
Smokeless tobacco	96
Pipe smoker	23

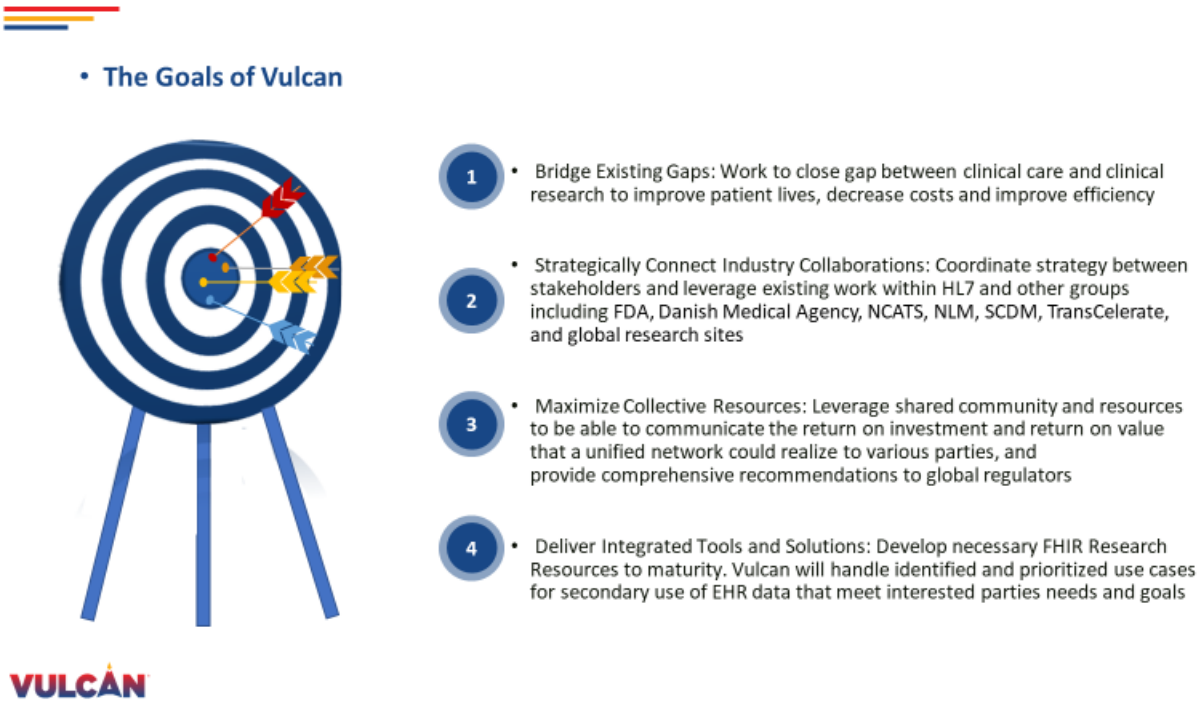
Dr. Bertagnolli, along with ASCO, ASTRO, Alliance for Clinical trials in Oncology Foundation, Society of Surgical Oncology, MITRE and FDA, are therefore working on the developing standard computable data formats known as Minimal Common Oncology Data Elements (mCODE) to achieve data interoperability and enable progress in clinical care quality initiatives, clinical research and healthcare policy development. They have launched CODEX (Common Oncology Data Elements eXtensions, http://hl7.org/CodeX) as an HL7 FHIR Accelerator: a community and platform to accelerate interoperable data modeling and implementation around mCODE, thus leading to step-change improvements in cancer care and research. A diagram showing mCODE follows.



mCODE can support various use cases, including clinical trials data management, clinical trials matching, registries, clinical care pathways and others.

- The introductory presentation of the **Vulcan Accelerator**, a new HL7 Accelerator, was provided next by **Amy Cramer of Pfizer.**

This Accelerator grew from the increasingly digital healthcare environment. An initial diverse group convened in September 2019 as "Experts Blazing the Path Forward for Research". HL7 initiated the FHIR Accelerator Program "to motivate and support market collaborations seeking to accelerate the availability of FHIR to tackle important interoperability needs". The Vulcan Accelerator that has now been launched includes technology vendors, research academia, patients, government agencies, industry groups, technology vendors and standards development organizations. The goals of the Vulcan Accelerator are shown below.



Ms. Cramer elaborated on what the Vulcan Accelerator is:

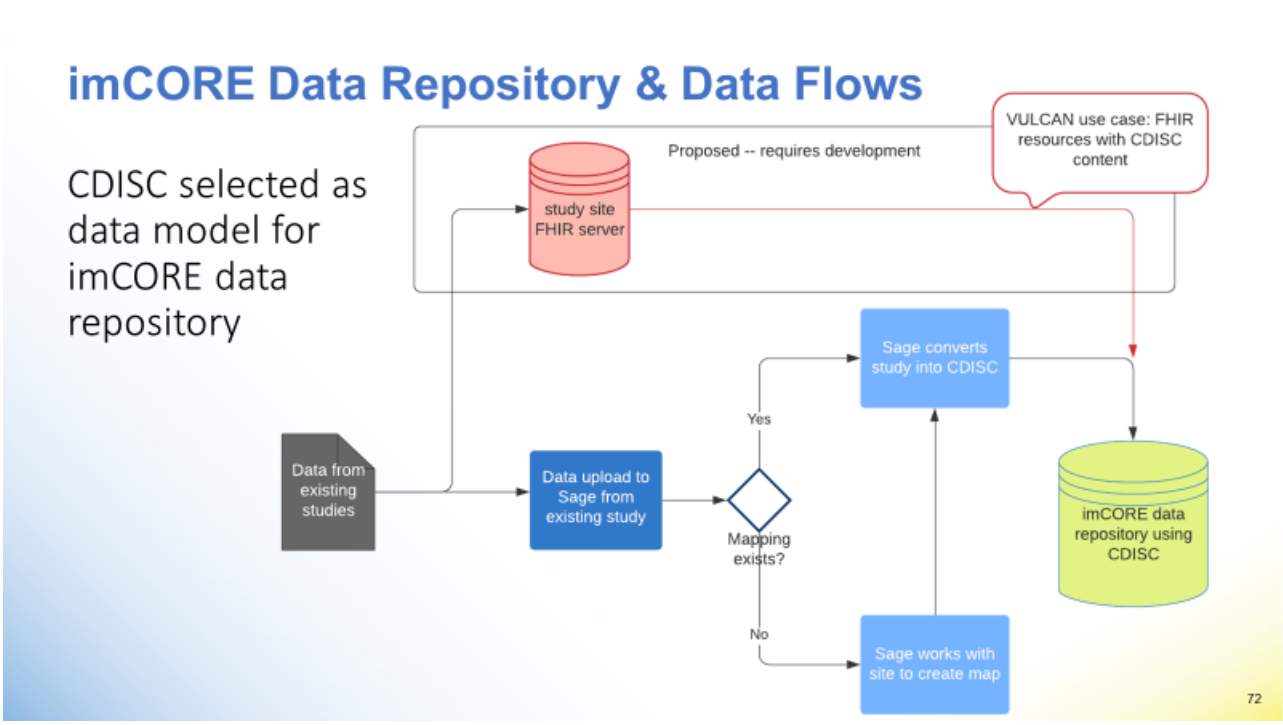
- focused on data mapping and standards for particular use cases that connect healthcare with clinical and translational research
- a collaborative membership-based project group with defined membership operated under HL7
- modeled after other accelerators, but ground-breaking
- initiated by key organizations, including TransCelerate, FDA, NIH/NCATS, SCDM,NIH/ NLM and academic research sites

The organizational structure includes an Advisory Council, a Steering Committee, an Operations Committee and Project Teams. Potential use cases are identified by the Operations Committee; these are then prioritized by the Steering Committee. The initial use cases will be: phenopackets in EHR, SoA, and RWD.

- A presentation on **Unleashing the Full Potential of Data to Accelerate Breakthroughs in Cancer Immunotherapy for Patients** by **Jonathan Chainey, Global Head of Data Standards and Governance and PD Biometrics at F. Hoffmann-La Roche & Genentech** brought together a number of the prior presentations in this seminar by focusing on an ongoing research initiative to treat cancer patients.

Mr. Chainey introduced imCORE (**Im**munotherapy **C**enters **o**f **R**esearch **E**xcellence), which is a global network on 27 leading cancer immunotherapy research sites. This project is funded by Roche and co-led with Sage Bionetworks. The goals are to advance and accelerate cancer immunotherapy R&D and harmonize clinical and biomarker data to support data sharing. There are currently over 50 investigator-initiated research studies. The locations are 33% within the US and 67% ex-US and they include various types of studies - clinical (33%), translational (32%, pre-clinical (28%) and data mining (7%).

When surveyed for the common data model that was being used at these sites, 55% responded that they were using "a locally-developed model" while 5% used CDISC and 5% used mCODE. The data repository and data flows that are being used for this project are shown in the following slide.



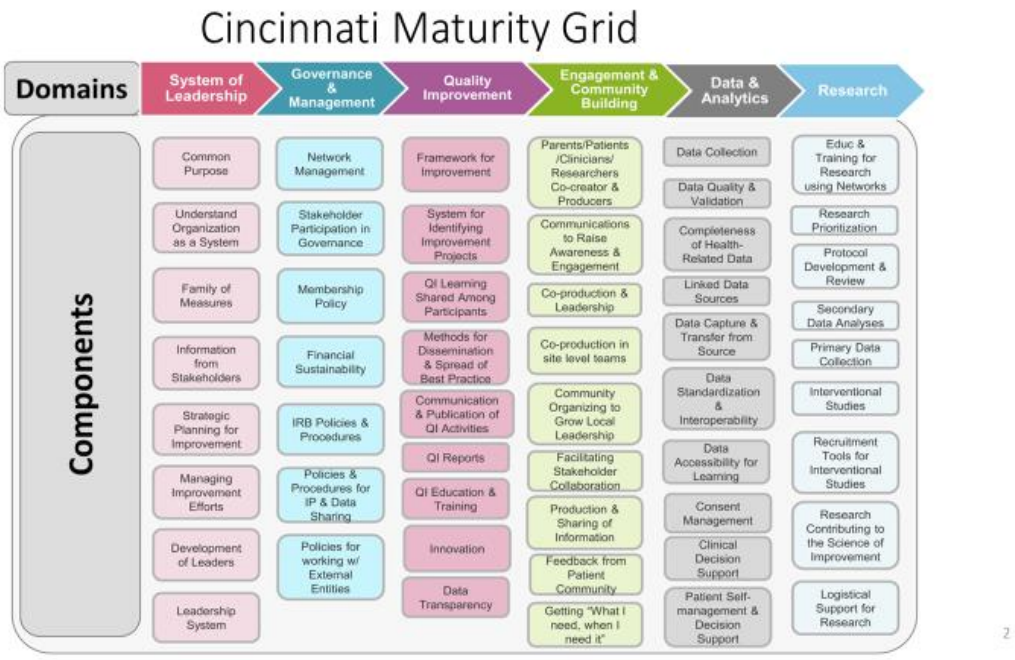
Mr. Chainey closed the presentation by describing the 'bigger picture" in which healthcare represents the largest 'circle of data'; academic research has a much smaller proportion while clinical research has the smallest, in comparison. With respect to healthcare data, SNOMED CT, ICD codes, HL7 and mCODE are used as standards/terminology/codelists while clinical research data are submitted to FDA in CDISC format. Mr. Chainey stated that it is "time to align"! Next steps for imCORE will be to build the data repository using the data model from Roche (i.e. CDISC and extensions), partner with imCORE research sites to increase data capabilities, especially with respect to CDISC and HL7 FHIR and then continue to seek alignment through Vulcan and FDA Common Data Model Harmonization using BRIDG.

Seminar 3: Expanding Learning Through Systems, Standards and Research

- The first presenter in this third eSource Symposium seminar on Learning Through Systems, Standards and Research was **Dr. Charles Friedman, Head of the University of Michigan Learning Health Sciences Department**. His presentation was on "**A Maturity Model for Learning Health Systems".**

Dr. Friedman introduced Learning Health Systems by stating two frequently asked and related questions: "How LHS is my LHS?" and "Operationally (not conceptually), what exactly is an LHS?" Capability Maturity Models (CMMs) point the way to these answers, the first being that CMMs create scales against which progress can be measured and the second one is that the zenith of CMM operationally defines a 'complete' LHS.

Pioneering work in the area of CMMs has been done in Cincinnati and was published in the Learning Health Systems Journal (Lannon, C. Schuler, C., Seid, M., Margolis, P. et al, 26 June 2020). The maturity grid they have published is shown in the slide below.



This is an example of the components and domains identified by Cincinnati for their learning network, i.e. a network of organizations. However, LHSs exist at differing levels of scale and CMMs are scale-specific. LHS levels can be for a single organization, for a network of organizations for a state/territory/region or for a nation or a planet. The Learning Health Community is currently working on a CMM that would pertain to a single organization.

The LHC approach focuses on such organizations as 'learning entities'. The 15-member working group that is developing this CMM was formed in October 2019 and has identified four pillars of the model and multiple maturity indicators within each pillar. These pillars consist of the following:

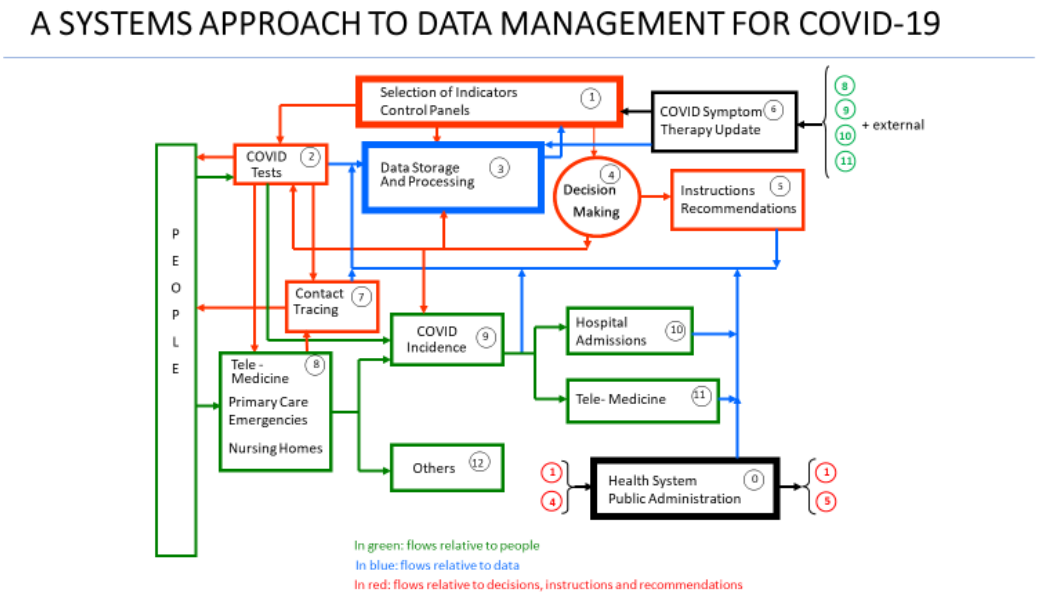
- **Governance and Leadership**: capability to set priorities and allocate resources
- **Execution**: capability to carry out complete improvement cycles
- **Infrastructure**: capability to support multiple simultaneous improvement cycles with shared socio-technical services
- **Culture and Values**: capability to create/define/act upon shared beliefs (e.g. equity, collaboration, and inclusion) that support other capabilities/pillars

Dr. Friedman closed by describing the next steps for this volunteer LHC working group that is developing the initial draft (v1.0) of the CMM, which will then be reviewed with a larger group to discuss and refine to produce v2.0. This version will be disseminated for open comment, which should enable the development of v3.0. Formal endorsement will be sought from a range of stakeholders.

❖ The second presentation for the 27 August Seminar was **by Dr. Franciso Ros** who is the **Former Secretary of State for Telecom and the Information Society in the Government of Spain** who spoke about **"A Globally Applicable Data-Driven Systems Approach to Covid-19".**

Critical groups for the spread of infections such as Covid-19 include the elderly (e.g. nursing home residents), chronic disease patients, health professionals, temporary workers, and socially fragile communities. Unfortunately, there has been a lack of good data to appropriately deal with these groups and other vulnerable groups. Dr. Ros also posed the questions of whether there has been diverse and partial cognitive bias in the decisions made from the interpretation of disease signs that are generated by a multilayered pandemic. "Have we been managing a Digital Era pandemic with Analog Era data and handicraft procedures?"

Dr. Ros proceeded to propose a Systems Approach to Data Management for the pandemic, showing a diagram with multiple boxes that represent various data sources along with indicators that fall into three main categories: clinic-related, territory-related and health system-related. Each of these indicator areas is associated with a number of metrics from data that should be collected and shared. For example, clinic-related indicators derive from such data as a) the number of cases by age group requiring hospitalization/ICU; b) the rate of new cases among critical groups and c) the number of closed cases by resolution (cured or deceased). The diagram is shown below.



The data input for each of the boxes in the diagram should be:

- Accurate and meaningful
- Consistently defined
- Represented in a standard format

Johns Hopkins Coronavirus Resource Center has been the source of much of the data on this pandemic; they report on their website that the data are from multiple sources and are not always consistently released or updated. The New York Times reported on 13 July 2020 that health officials are dealing with a "broken data system that sends incomplete results in formats they can't easily use".

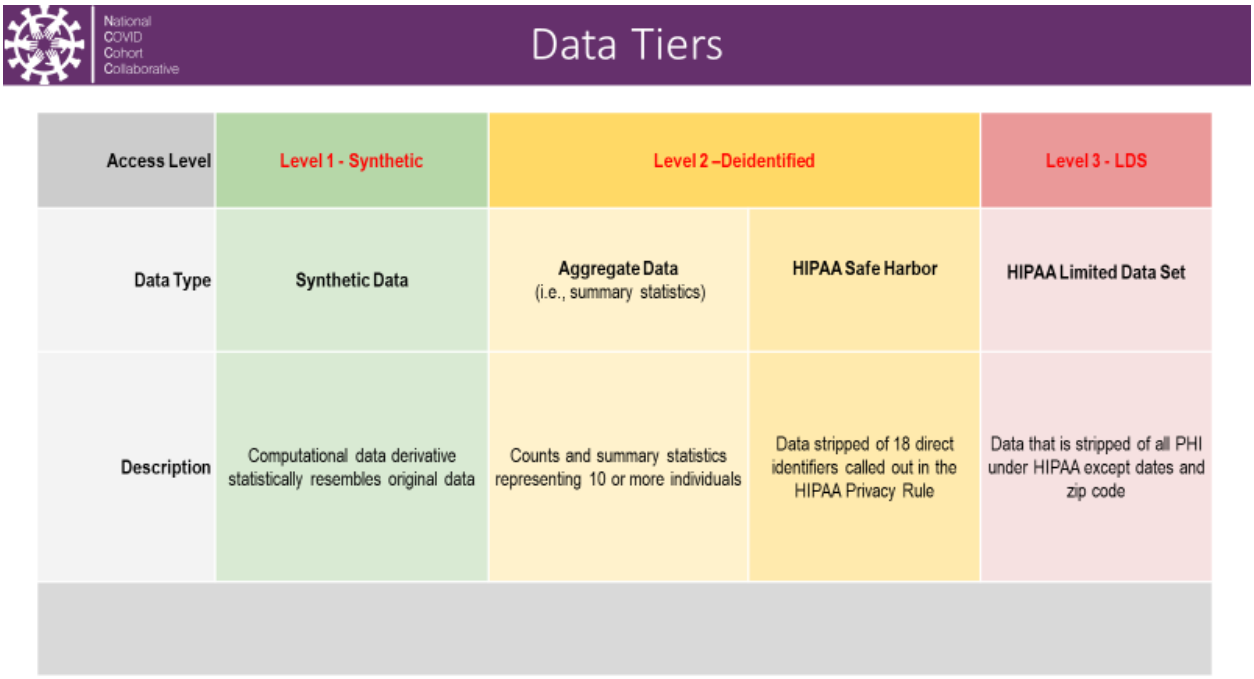
Global data standards now exist for collection of data from clinical research on new and existing therapies and vaccines to fight Covid-19. Such standards also exist for reporting on adverse events and registering new clinical research trials. Common data models (CDMs) exist; however, they are not well-aligned. EHRs that capture health data remain quite disparate. Therefore, areas of opportunity for developing global consensus-based standards to support a complete systems approach for Covid-19 are in the reporting of test results, reporting health data from EHRs and contact tracing and mobility. Standards are most valuable when they are widely adopted and consistently implemented.

- **Dr. Ken Gersing of the NIH National Center for Accelerating Translational Science (NCATS)** presented next on the topic of the **"National COVID Cohort Collaborative (N3C)".**

To date, the mode of operation of the various research-related networks in the United States has been to send a federated query to an academic research organization (ARO), which returns aggregated results. These data are modeled differently depending on the network, i.e. PCORNet common data model (CDM) and the i2b2/ACT CDM. Sentinel, which was created to provide safety information for the FDA has yet another CDM and the Observational Health Data Sciences and Informatics program (ODHSI) OMOP CDM. This means that not only does each ARO need to support varied models depending on which research networks they wish to work with, but the data reside at the ARO. In contrast, a model with centralized analytics would have data that resides securely in a central enclave.

In an interest to harmonize the four CDMs and reduce the burden on AROs, these CDMs have been mapped to a central information model, which was described by Denise Warzel of NIH/NCI during the 25 August seminar. This central model is called the Biomedical Research Integrated Domain Group (BRIDG) model, which is an HL7, CDISC and ISO standard. Ultimately, FDA, NCI and NCATS, are working together on CDM harmonization leveraging the FHIR USCORE, which was described by Dr. Ida Sim in the 25 August seminar.

In the interim, since all of this harmonization is still in progress and FHIR is not yet widely implemented, the NCATS/N3C project has designed a data pipeline that can accommodate the various CDMs, harmonize and transforming the data into OMOP. Collaborative analytics can then be done on the data in this format. All participants in the N3C project must sign a Data Use Agreement that provides privacy protection while promoting broad access. This agreement pertains only to Covid-related data and states that there will be no re-identification of the data source or individuals and no download of raw data. The analytics platform will be open to all researchers. The various data tiers are shown in the slide below.



At the time of this webinar (27 August 2020), data was available from 6 sites (over 16,000 Covid-19 patients out of a total of nearly 300,000 patients). Data use agreements were still being signed to access data from additional NCATS sites. This is an ongoing project to test this current system, which will enable learning and the design of a new system that will hopefully leverage HL7 FHIR.

- **Dr. Sam Volchenboum of the University of Chicago** spoke next about their work on international data sharing and the Pediatric Cancer Data Commons (PCDC). His presentation was entitled **"Transforming the Way Researchers Share Data."**

Dr. Volchenboum started with a historical perspective around the move from paper to electronic data capture, showing a clinical research site with numerous binders on the shelves and multiple laptops (one for each different study). He also showed a slide that reinforced the point made on 25 August by Dr. Bertignolli---that databases often have numerous ways in which the same data element is expressed. Dr. Volchenboum used age as an example, showing 30 ways this data point was expressed in the Gene Expression Omnibus (GEO).

Dr. Volchenboum then discussed essential elements for building a commons, specifically:

- Consortium – building trust between groups
- Data Governance – publication policy, approving data contribution and use
- Scientific Goals – why build it? What data to include?
- Data Dictionary – consensus data elements
- Data Transformation and Aggregation – into the consensus standard, by statisticians and data scientists
- Funding – building and sustaining the commons

The PCDC operates with a principle of consensus-based decision making and data sharing, and this is the foundation of creating a consensus data dictionary. The PCDC provides updates and works with the NCI,

leveraging the caDSR and NCI Terminology (NCIt). They are building a pan-pediatric cancer data dictionary through systematic reviews of existing data dictionaries, development of a PCDC model, and generating a template for future data dictionaries.

Dr. Volchenboum showed examples of the PCDC data model variables. He stated that choosing the right standard is important. Considerations include consistency across cancer groups, interoperability with non-PCDC data groups (e.g., St. Jude Children's Research Hospital and the Dana-Farber Cancer Institute) and the ability to connect clinical data to outside data sources such as genomics and imaging. The PCDC works with the NCI for standardization; the PCDC Data Dictionary Work Group is composed of international pediatric oncologists, statisticians, and data standards experts. Unfortunately, as others during the eSource Symposium Seminars have indicated, there are too many different standards that are currently in play. These include Health Level Seven (HL7) with FHIR and mCODE; CDISC with SDTM; OHDSI with the OMOP Common Data Model; PCORI with the PCORNet Common Data Model; Sentinel and I2b2 with their Common Data Models; and BRIDG, which is an ISO/HL7/CDISC Information Model. And, pediatric data standards have not been fully developed in any case.

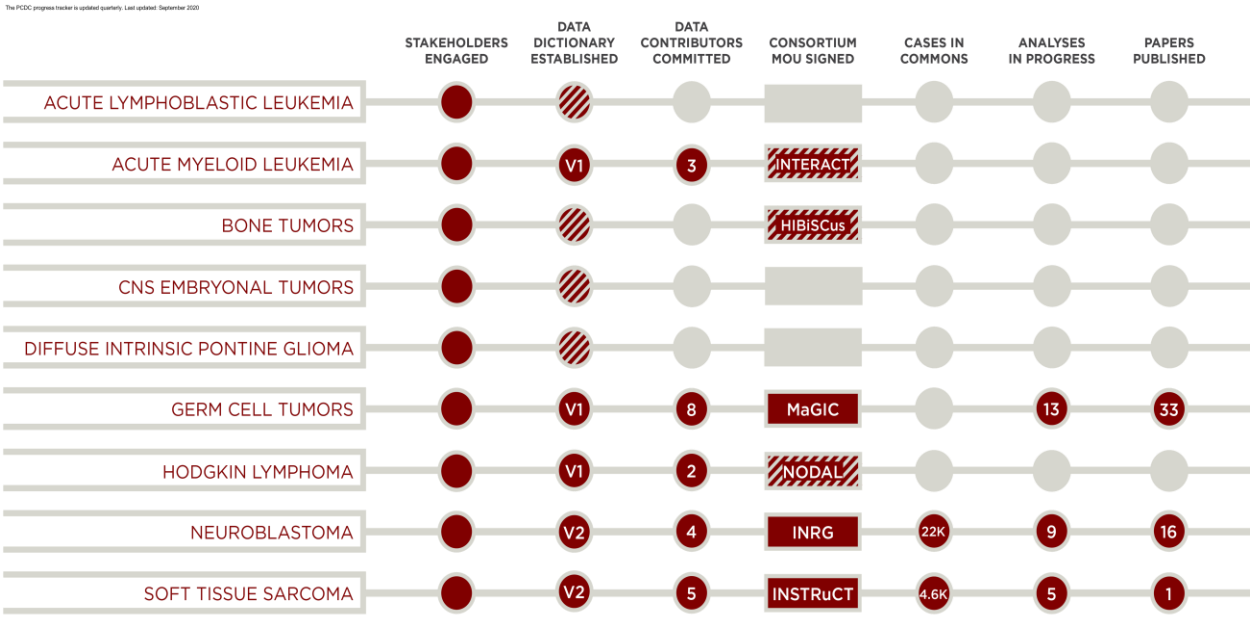
Dr. Volchenboum then discussed the value of a Commons, such as PCDC. Particularly since pediatric cancer occurs quite rarely, sharing data is important and valuable. Commons can enable a number of different types of research:

- Correlating biomarkers with clinical outcomes across trials
- Understanding impact of dose modifications across trials
- Performing patterns of failure analyses
- Examining toxicity prognosticators
- Validating consensus staging definitions across trials
- Validating prognostic scoring systems
- Enhancing risk stratification
- Prognosis of rare subgroups
- Age-related differences in therapeutic response
- Disparities analyses

Guiding Principles governing the PCDC are:

- Their goal, which is to lift barriers and connect researchers to data
- Stakeholder approval for data release from any disease commons
- Contributor approval for data releasee from the original contributing group
- Representation on the Executive Committee from every disease group
- Recognition of regional differences – European and US legal regulations are not the same

Dr. Volchenboum spoke about the integrated pediatric cancer data commons, with clinical data from multiple different types of pediatric cancers from a number of different sources around the world. He showed a slide with the PCDC progress to date (which can be followed on their University of Chicago website http://commons.uchicago.edu):



Dr. Volchenboum closed by discussing their collaborative work through the Center for Cancer Data Harmonization (CCDH) within the NCI Cancer Research Data Commons (CRDC) Ecosystem. The CCDH is working to

a) **facilitate** retrospective and prospective semantic harmonization of data across nodes of the CRDC
b) **coordinate** the community to ensure implementation of standards that will facilitate interoperability of heterogeneous data types and CRDC resources
c) **find agreement** across the communities built around CRDC
    - match and extend data models
    - annotation, harmonization
    - quality assurance

Take home points from Dr. Volchenboum's presentation were:

- Studying pediatric cancer requires **collaboration and sharing**
- Data sharing needs to be built on a foundation of **trust and consensus**
- **Connecting disparate data types** and sources enriches research
- **Consensus data standards** are important for the success of national and international data ecosystems – allowing aggregation across trials and diseases
- Early adoption of data standards and consideration for the **lifecycle of the data** is critical to accelerating progress in discovery

- **Susan Winckler, RPh, Esq, FAPhA leads the Reagan-Udall Foundation for the FDA.** She gave a presentation on the recent work of this Foundation in collaboration with FDA and Friends of Cancer Research about a **"Covid-19 Evidence Accelerator – Experience in Use of RWE for Regulatory Decisions".**

The Foundation/Friends Covid-19 Evidence Accelerator (https://evidenceaccelerator.org/) is an initiative launched to provide a unique venue for major data organizations, government and academic resarchers, and health systems to gather and design quick-turnaround queries. The country's experts in health data aggregation and analytics are unified in an effort to quickly share and compare results and answer questions about treatments and other questions related to Covid-19, building evidence through the use of real world data (RWD).

"The Evidence Accelerator provides a 'safe space' for key players across the ecosystem to lead, scrutinize and 'get this right'." Priorities include data selection, protocol design, transparency, data provenance, data quality, analytic integrity, peer review and press interactions.

Susan Winckler presented the slide below to describe how the Foundation/Friends Evidence Accelerator, working with the FDA, adhere to a set of principles such that they CREATE and LEAD.

❖ The final presentation for the 27 August seminar was co-presented by **Dr. Martin Kohn of MedPreixAI, LLC and Wake Forest and Dr. Mary Tobin of ACRES.** Their presentation was on **"Systems Thinking and Synergy".**

Drs. Kohn and Tobin began this presentation by showing a very complex picture entitled 'System Overload' and then contrasted this diagram that was virtually impossible to comprehend with another diagram on the Air Transportation System, indicating that this is a 'functioning system at work'. Maintaining Global Air Transportation Safety requires a systems approach. The following slide compares statistics between drug safety and air transportation safety.

Healthcare is a very Complex System. Compared to systems that are Simple (easy to manage with existing methods) and Complicated (difficult, but amenable to rules and procedures), a Complex System is very difficult, with too many unknowns and unpredictable interactions. A possible solution for this Complex System of Healthcare was proposed:

- Empirical observation to quantify components of your model
- AI to identify patterns for study and inferences
- Progressively include more data types to approach personalization

A Health Systems Map was shown as an example. Applying Systems Thinking and Personalized Decisions to healthcare, one must study the influence of all components of the system model on the outcome to better manage patients with multiple chronic disease, define longer term outcomes (e.g. a healthier patient in five years) using decision support as a predictor.

This discussion on Systems Thinking Approach to Personalized Healthcare will be continued in 2021, hopefully at an in-person 'Think Tank' meeting, depending on how and/or when we emerge from the Covid-19 pandemic that has seriously altered 2020 in-person meeting plans.